The Qualitative Report Volume 9 Number 4 December 2004 770-792
http://www.nova.edu/ssss/QR/QR9-4/onwuegbuzie.pdf

Enhancing the Interpretation of "Significant" Findings: The Role of Mixed Methods Research

Anthony J. Onwuegbuzie
University of South Florida, Sarasota, Florida

Nancy L. Leech
University of Colorado at Denver and Health Sciences Center, Denver, Colorado

The present essay outlines how mixed methods research can be used to enhance the interpretation of significant findings. First, we define what we mean by significance in educational evaluation research. With regard to quantitative-based research, we define the four types of significance: statistical significance, practical significance, clinical significance, and economic significance. With respect to qualitative-based research, we define a significant finding as one that has meaning or representation. Second, we describe limitations of each of these types of significance. Finally, we illustrate how conducting mixed methods analyses can be used to enhance the interpretation of significant findings in both quantitative and qualitative educational evaluation and policy research. Consequently, mixed methods research represents the real "gold standard" for studying phenomena. Key Words: Quantitative Research, Qualitative Research, Mixed Methods, Significance, Meaning and Verstehen

Setting the Scene

One argument posited by proponents of mixed methods studies is that they address much more comprehensive research purposes than do quantitative or qualitative research alone (Newman, Ridenour, Newman, & DeMarco, 2003). Consistent with this assertion, and expanding on Rossman and Wilson's (1985) work, Greene, Caracelli, and Graham (1989) categorized the following five general purposes of mixed-methodological studies: (a) triangulation (i.e., seeking convergence and corroboration of findings from different methods that study the same phenomenon); (b) complementarity (i.e., seeking elaboration, illustration, enhancement, and clarification of the findings from one method with results from the other method); (c) development (i.e., using the findings from one method to help inform the other method); (d) initiation (i.e., discovering paradoxes and contradictions that lead to a re-framing of the research question); and (e) expansion (i.e., seeking to expand the breadth and range of inquiry by using different methods for different inquiry components). As observed by Greene et al. (1989), every mixed methodological study can be classified as having one or more of these five purposes.

In recent years, the advantages of mixed methods research have been increasingly recognized. In particular, as noted by Onwuegbuzie and Leech (in press), combining quantitative and qualitative research enables evaluation researchers to be more flexible and holistic in their investigative techniques, as they endeavor to address a range

of complex research questions that arise. Further, mixed methods research helps investigators to develop a conceptual framework, to validate quantitative results by linking the information extracted from the qualitative phase of the study, and to construct indices from qualitative data that can be utilized to analyze quantitative data (Madey, 1982). Also, by conducting mixed methods studies, researchers are in a better position to combine empirical precision with descriptive precision (Onwuegbuzie, 2003a). In addition, by employing a pragmatist lens (i.e., using both quantitative and qualitative techniques), rather than using a single lens (i.e., conducting monomethod studies), investigators are able to *zoom in* to microscopic detail or to *zoom out* to indefinite scope (Willems & Raush, 1969). As such, mixed research investigations afford researchers with the opportunity to combine macro and micro levels of a study (Onwuegbuzie & Leech, in press).

Compared to their monomethod counterparts, mixed methods researchers are more able to utilize quantitative research to inform the qualitative portion of research studies, and vice versa. For example, the inclusion of qualitative data can help investigators to explain relationships emerging from quantitative data. Similarly, the inclusion of quantitative data can help compensate for the fact that qualitative data typically cannot be generalized (Onwuegbuzie & Johnson, 2004). As such, mixed methods optimally involve the combining of methods that have complementary strengths and non-overlapping weaknesses; this is known as the *fundamental principle of mixed methods research* (Johnson & Turner, 2003). Indeed, because mixed methods research involves combining quantitative and qualitative approaches in some manner within the same inquiry, investigators using this paradigm are able to probe further into a dataset to understand its meaning and to use one method to verify findings stemming from the other method (Onwuegbuzie & Teddlie, 2003).

However, an even more important rationale exists for conducting mixed methods research that has received little or no attention. Specifically, we believe that this class of research can be used to enhance the interpretation of *significant* findings in educational evaluations. Thus, the goal of this present essay is to outline how this can be accomplished. First, we define what we mean by significance in evaluation research. With regard to quantitative-based evaluations, we define the four types of significance, as identified by Leech and Onwuegbuzie (in press): statistical significance, practical significance, clinical significance, and economic significance. With respect to qualitative-based evaluations, we define a significant finding as one that has meaning or representation. Second, we describe some of the limitations of each of these types of significance. In particular, we contend that all of these types of significance provide partial information at best, and that sole reliance on any of these indices can lead to misleading interpretations of the data, which, in turn, could adversely affect ensuing policies.

Finally, we illustrate how conducting mixed methods analyses can be used to enhance the interpretation of significant findings in both quantitative- and qualitative-based evaluation research. In terms of quantitative-based research, we demonstrate how the collection, analysis, and interpretation of qualitative data can aid the interpretation of statistically significant, practically significant, clinically significant, and economically significant findings. With respect to qualitative findings, we outline how quantitative data collection, analysis, and interpretation can add meaning.

Significance in Educational Research

Significance in Quantitative Research

Thompson (2002) identified the following three types of significance in quantitative research: statistical significance, practical significance, and clinical significance. Each of these types of significance is represented by an array of indices from which criteria can be used to determine the level of significance. These types of significance are each discussed below.

Statistical significance

Statistical significance indices (i.e., *p* values) estimate the probability that results from the sample could have occurred if the null hypothesis is true (Cohen, 1997). As noted by Cohen (1988), if the null hypothesis is true, the probability of the sample result is no greater than $\forall$, the level of significance that is set in advance by the researcher (i.e., *a priori*). If $\forall$ is small (e.g., .05), the researcher would reject the null hypothesis at the *a priori* level of significance. Conversely, if the *p* value is greater than or equal to this level, then the researcher concludes that the null hypothesis cannot be rejected at the $\forall$ level of significance (Cohen, 1988). Thus, the null hypothesis significance tests are conducted to determine whether an observed result is due to chance.

Practical significance

Practical significance represents the educational value of the results (Gay & Airasian, 2003). In other words, the practical utility of a result can be improved by reporting practical significance. The most common way of assessing the practical significance of a finding is via the use of effect sizes. An effect size represents a term given to a family of indices that measure the size of a difference or relationship (Onwuegbuzie, Levin, & Leech, 2003). Moreover, an effect size index provides information about the theoretical or applied significance of a result (Thompson, 2002; Vacha-Haase, 2001; Vaske, Gliner, & Morgan, 2002). According to Cohen (1988):

> Without intending any necessary implication of causality, it is convenient to use the phrase 'effect size' to mean 'the *degree* to which the phenomenon is present in the population,' or 'the degree to which the null hypothesis is false.' By the above route it can now readily be clear that when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific non-zero value in the population*. The larger this value, the greater the *degree* to which the phenomenon under study is manifested. (pp. 9-10)

Kirk (1996) classified 61 different effect-size indices. More recently, Huberty and his colleagues (e.g., Huberty & Lowman, 2000) developed new effect size indices that they refer to as Group Overlap indices. All of these indices can be classified into two

broad categories: (a) variance-accounted-for measures, also known as "r^2 family" effect-size indices (e.g., r^2, R^2, η^2, ω^2) and (b) measures of standardized differences, also known as "*d* family" effect-size indices (e.g., Cohen's *d*, Glass's Δ, Hedges' *g*) (cf. Majova-Seane, 2003). Additionally, effect-size indices also can be classified as being "uncorrected" or "corrected" indices (Kirk, 1996; Olejnik & Algina, 2000). For example, in multiple regression, researchers can compute and interpret R^2 (uncorrected effect size) and/or adjusted R^2 (corrected effect size) (Cohen, 1988).

Clinical significance

As defined by Kazdin (1999), clinical significance represents the extent to which an intervention makes a real difference to the quality of life of the participants or to those with whom they interact or encounter. More specifically, Kendall, Marrs-Garcia, Nath, and Sheldrick (1999) refer to clinical significance as the

> convincingness or the amount of change linked to treatment…when one is interested in clinical significance, two questions arise: (a) Is the amount of change that has occurred, presumably because of treatment, large enough to be considered meaningful and (b) are treated individuals distinguishable from normal individuals with respect to their primary complaints following treatment? (p. 285)

According to Vacha-Haase (2001), the overall goal of clinical significance is to "report data from research that can be utilized by consumers, that is, clinicians providing direct services" (p. 15). Further, Leech and Onwuegbuzie (in press) discussed two approaches to understanding clinical significance: (a) the reliable change index that represents the amount of change and (b) the normative comparisons that represent how distinguishable the individual is from a normative sample. Kazdin posits that it is possible for interventions that yield no statistically significant or practically significant effect to be clinically significant.

Economic significance

Recently, Leech and Onwuegbuzie (in press) identified a fourth measure of significance in quantitative research, namely, what they termed *economic significance*. Leech and Onwuegbuzie defined economic significance as the cost-effectiveness ratio pertaining to an observed finding. More specifically, economic significance refers to the economic value of the effect of the intervention. The major advantage that measures of economic significance have over the other three types of significance used in quantitative research (i.e., statistical significance, practical significance, and clinical significance) is that they incorporate both the effects and costs of educational choices, treatments, or programs. Leech and Onwuegbuzie noted that "if an intervention prevents a child from dropping out of school, then the economic significance could represent that child's economic contribution to society, weighted by the probability of selected risk factors (e.g., probability of no future incarceration)" (p. 5).

At least five classes of economic significance indices (ESIs) have been identified: cost effectiveness, cost benefit, cost utility, cost-feasibility, and cost sensitivity. A cost-

effectiveness ESI provides information about the cost per level of effectiveness or the maximum effectiveness of an intervention per level of cost. Cost-benefit ESIs yield estimates that compare costs and benefits, determining the amount of benefit in relationship to the cost. Cost-utility ESIs provide information about the cost of the interventions relative to their estimated utility or value of their outcomes. Cost-feasibility ESIs yield information exclusively about the cost of an intervention in order to determine whether it is within the boundaries of the budget or other available resources. These indices are calculated by comparing the cost of an intervention with the given budget for an intervention. Finally, cost-sensitivity ESIs represent estimates of economic significance that build in uncertainty into the estimate of effectiveness, cost, benefit, utility, and/or feasibility (Leech & Onwuegbuzie, in press; Levin & McEwan, 2001).

Significance in Qualitative Research

The goal of qualitative research typically is to obtain insights into particular educational, social, and familial processes and practices that exist within a specific location (Connolly, 1998). Bogdan and Biklen (2003) state one of the features of qualitative research is to define "how people negotiate meaning" (p. 6). In an attempt to gain insights, qualitative researchers tend to seek to extract meaning from their data. That is, qualitative researchers study phenomena in their natural settings and strive to make sense of, or to interpret them with respect to the meanings people bring to them (Denzin & Lincoln, 2000). Schwandt (2001), in the *Dictionary of Qualitative Inquiry*, defines "meaning" as, "a taken-for-granted assumption in qualitative inquiry that it studies meaningful social *action*…it cannot be adequately described in purely physical terms" (p. 153). He goes on to state that a physical action has different meaning for different people, and, "the significance of the action cannot be adequately explained in terms of a behaviorist stimulus-response model" (p. 153). Thus, as noted earlier, a significant finding in qualitative research is one that has meaning or representation.

According to interpretivists, what differentiates observations from human beings (i.e., data pertaining to the social and behavioral sciences) from observations stemming from physical objects (i.e., data from the physical sciences) is that the former is fundamentally meaningful (Schwandt, 2000). For a particular human behavior to be understood, it must have a specific intentional content that indicates the type of behavior it is, or that what a behavior means can be understood only with respect to the system of meanings to which it belongs, or both (Outhwaite, 1975). Therefore, for interpretivists, in order for a behavior to be understood, the meaning that underlies that behavior must be understood—that is, *Verstehen* must be achieved. Moreover, interpretivists contend that the subjective meaning of action can be understood in an objective manner (Schwandt, 2000).

Carspecken and Apple (1992) believe there are three steps involved in the "analysis of meaning" (p. 519). These steps include: (a) noting possible meanings in field notes, (b) reconstructing normative factors [they state that "[u]nderstanding meaning, then, involves taking first-, second-, and third-person positions with respect to an act and this can be done only with reference to certain norms assumed to be in play" (p. 519)], and (c) subjective states of the individuals must be reconstructed. These steps help the researcher find meaning within the data.

For applied ethnographers, cultural meanings that stem from the interactions of groups are of particular interest. Also of interest to ethnographers is to study how "cultural meanings might be exchanged and negotiated as a result of intracultural attempts to find solutions to problems" (Chambers, 2000, p. 856). According to Chambers, modern day ethnographers tend to focus on how "people fashion culturally meaningful expressions from fields of experience in which meaning is routinely contested, and where culture is perennially under construction" (p. 857).

Unlike quantitative research articles, which primarily can be interpreted through their results sections, tables, figures, and graphs, qualitative research articles carry their meaning throughout the entire text. Thus, its meaning is in its reading (Richardson, 2000). In turn, significance is extracted from its reading.

Limitations of Significance Indices Used in Educational Research

Quantitative Research: Limitations of Significance Indices

Statistical significance

Although *p* values help researchers to rule in or to rule out chance as an explanation for an observed finding, statistical significance testing has several serious limitations. In particular, statistical significance testing is often subjected to misunderstanding, abuse, and misuse stemming from the fact that many researchers do not understand the logic of what statistical significance tests do (Cohen, 1997), and, thus, they tend to interpret *p* values incorrectly (Thompson, 2002). For instance, some researchers believe that statistical significance indicates whether a result is true for a population, as well as indicates the strength or size of an effect (e.g., they believe that a *p* value < .05 is less important or significant than a *p* value < .001). An additional limitation of statistical significance testing is that all *p* values represent a function of the underlying sample size. That is, holding everything else constant, the smaller the sample, the smaller the probability of obtaining a statistically significant result (Fan, 2001; Kirk, 1996; Thompson, 1993). Further, as stated by Kirk (1996), statistical significance testing "turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision" (p. 748), with this dichotomous decision process often leading to "the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions" (Kirk, 1996, p. 748) because of minor discrepancies in the size of the samples or another aspect of their designs. However, according to Leech and Onwuegbuzie (in press), the most serious limitation of statistical significance testing is that "not only do most researchers not understand what information can be found through statistical significance testing, but policy makers and change agents are usually unable to glean helpful information from a reported *p* value of .05" (p. 8).

Practical significance

Although indices of practical significance, such as effect-size measures, provide useful information about the magnitude of an effect or a relationship, they also have numerous important limitations. In particular, Onwuegbuzie and Levin (2003) identified

nine limitations concerning measures of effect size: (a) effect sizes can vary as a function of the investigator's research objective (i.e., theory application or effects application); (b) effect sizes can vary as a function of the investigator's research design and experimental conditions; (c) researchers can select from a variety of effect-size measures to justify different (possibly self-serving) points; (d) guidelines for interpreting effect size magnitudes are generally inconsistent and arbitrary; (e) effect sizes can vary as a function of sample size and sample variability; (f) effect sizes can vary as a function of the variability of the outcome measure (both between and within samples); (g) effect sizes are sensitive to departures from normality; (h) effect sizes can vary as a function of the score reliability of the outcome measure; and (i) effect sizes can vary as a function of the scale of measurement used (i.e., nominal, ordinal, interval, ratio).

Further, measures of effect size are not always meaningful or useful for consumers of research (Onwuegbuzie, Levin, & Leech, 2003). As noted by Leech and Onwuegbuzie (in press), "[I]t is difficult for a policymaker to know how to act upon a Cohen's *d* effect size of .28, even if confidence limits are placed around this value" (p. 9). Thus, effect sizes, per se, do not help stakeholders and policymakers to select the most effective intervention.

Clinical significance

Although measures of clinical significance are useful in clinical studies, these indices have limitations. The primary limitation is that clinical significance often is only relevant in clinical settings. Therefore, unfortunately, it has limited applications in the educational context. Second, the clinical significance of results typically is difficult to ascertain because of its qualitative nature. According to Kendall et al. (1999), when normative comparisons are of interest, the first step is to define what normative is for the particular context. However, for many researchers and clinicians, this can be difficult. Furthermore, judgments about levels of clinical significance depend, at least in part, by the person making the interpretation (e.g., the researcher, evaluator, policymaker, stakeholder). Additionally, no simple formula exists for determining how much change a client/participant must experience for a judgment of clinical significance to be given. In fact, individuals responsible for making these judgments often disagree as to which criteria to use. Kazdin (1999) notes that clinical significance can be found "when symptoms change a lot; when they change a little; and when they do not change at all, but the client is better able to cope with them" (p. 333). This culminates in much ambiguity as to what exactly clinical significance is assessing.

Economic significance

Even though economic significance indices provide useful information for policymakers and consumers in units that are understandable to them, they have limitations. One limitation is difficulty with interpreting the results, especially if there are multiple measures of effectiveness. Further, as noted by Leech and Onwuegbuzie (in press), it is difficult in some situations to estimate the true cost. An additional weakness is that this measure should only be used when comparing two alternatives; it does not yield accurate information when only one choice is used.

Qualitative Research: Limitations of Significance Indices

Although the strength of qualitative research lies in its focus on extracting meaning, like quantitative research, it still has serious limitations. In particular, these limitations include researcher prejudice and bias, observer effects, and writing about qualitative research so that readers can replicate the study.

Due to the evaluator being the key instrument in qualitative research, the evaluator's prejudice and bias can be introduced into the findings and results of studies (LeCompte, 1987). Qualitative researchers are concerned with bias and prejudice. Thus, they attempt to "objectively study the subjective states of their subjects" (Bogdan & Biklen, 2003, p. 6). Yet, due to the evaluator being a key part of the collection, analysis, and interpretation of data, bias and prejudice will always be a concern and limitation.

A second limitation of qualitative research is observer effects. Observation is one of the most commonly used type of data collection in qualitative research. Deyhle, Hess, and LeCompte (1992) conclude that observer effects are an ethical issue for qualitative researchers. Observer effects need to be assessed for the "real impact of the researchers' presence…(both for methodological veracity and for sociological and psychological impact)" (pp. 615-616). Even though qualitative researchers discuss the issues involved in observation and strive to eliminate these by "using rigorous methods to validate observations" (Patton, 1990, p. 201), issues of observer effects still abound.

Many qualitative researchers believe that qualitative research should be separate from quantitative research; this separatism includes the underlying assumptions, methods, analysis, and writing about the research. Positivists state that due to this last difference of how the research is written in article form, that qualitative research is, "fiction, not science, and that these researchers have no way to verify their truth statements" (Denzin & Lincoln, 2000, p. 8). Thus, a major limitation of qualitative research is that many qualitative studies have not been written so that the methods are understood and so they might be replicated. As Guba (1981) stated, "the naturalistic approach is likely to be tarred with the brush of 'sloppy research' " (p. 90). This creates a major limitation in that many qualitative studies are not written in such a way that other researchers can ascertain the research design and choice of strategy of inquiry: What methods of data were collected? What did the data look like? How were the results determined? What steps were involved in reaching the conclusion? These components need to be clear and understandable in every article from a qualitative study. Recently, this issue has been given more thought (Anfara, Brown, & Mangione, 2002), but we believe more information is needed so that qualitative studies do not "remain private and unavailable for public inspection" (Constas, 1992, p. 254).

These limitations give rise to what are referred to as the triple crisis of representation, legitimation, and praxis (Denzin & Lincoln, 2000). These three crises make problematic two key assumptions of qualitative research. According to Denzin and Lincoln (2000):

> The first is that qualitative researchers can no longer directly capture lived experience. Such experience, it is argued, is created in the social text written by the researcher. This is the representational crisis. It confronts

> the inescapable problem of representation, but does so within a framework that makes the direct link between experience and text problematic.
>
> The second assumption makes problematic the traditional criteria for evaluating and interpreting qualitative research. This is the legitimation crisis. It involves a serious rethinking of such terms as *validity*, *generalizability*, and *reliability*, terms already retheorized in postpositivist…, constructivist-naturalistic…, feminist…, interpretive…, poststructural…, and critical…discourses. This crisis asks, "How are qualitative studies to be evaluated in the contemporary, poststructural moment? The first two crises shape the third, which asks, Is it possible to effect change in the world if society is only and always a text? Clearly these crises intersect and blur, as do the answers to the questions they generate… (p. 17)

The crises of representation, legitimation, and praxis threaten qualitative researchers' ability to extract meaning from their data. In particular, lack of representation means that the evaluator has not adequately captured the data. Lack of legitimation means that the extent to which the data have been captured has not been adequately assessed, or that any such assessment has not provided support for legitimation. Thus, the significance of findings in qualitative research is affected by these crises.

A Framework for Enhancing the Interpretation of Significant Findings

As can be seen from the preceding sections, the ability of both quantitative and qualitative evaluators to extract significance from their data is compromised by the limitations inherent in the method of extraction. Moreover, we contend that all of the types of significance associated with quantitative and qualitative research provide partial information at best, and that sole reliance on any of these indices can lead to misleading interpretations of the data. Thus, neither quantitative nor qualitative research, per se, is optimal in interpreting significant findings. As such, we advocate the use of mixed methods data analyses to enhance the interpretation of significant findings in both quantitative- and qualitative-based evaluation research. In terms of quantitative-based research, we contend that the collection, analysis, and interpretation of qualitative data aid the interpretation of statistically significant, practically significant, clinically significant, and economically significant findings. In terms of qualitative-based research, we assert that the collection, analysis, and interpretation of quantitative data add meaning and enhance the achievement of *Verstehen*. In what follows, we illustrate how this can be accomplished.

Enhancing the Interpretation of Significant Findings in Quantitative Research

Including the collection, analysis, and interpretation of qualitative data for the purpose of enhancing the interpretation of statistically significant, practically significant, clinically significant, and economically significant findings can be undertaken either

concurrently or sequentially. Indeed, an array of parallel, concurrent, and sequential mixed methods data analysis techniques can be employed to shed more light on significant findings emerging from quantitative data analyses.

Parallel mixed analysis

In order to conduct a parallel mixed analysis, the following three conditions should hold: (a) both sets of data analyses (i.e., quantitative and qualitative data analyses) should occur separately, (b) neither type of analysis builds on the other during the data analysis stage, and (c) the results from each type of analysis are neither compared nor consolidated until *both* sets of data analyses have been completed. Of the three mixed analysis techniques, parallel mixed analyses involve the least amount of mixing because mixing or integration does not occur until the data interpretation stage of the mixed methods research process, if at all. Nevertheless, parallel mixed analyses can still be utilized to enhance the interpretation of statistically significant, practically significant, clinically significant, and economically significant findings. For example, Onwuegbuzie (1997) investigated the anxiety experienced by graduate students from non-statistical disciplines, who wrote research proposals in an introductory-level research methodology course. For the quantitative portion of the research (i.e., Study 1), students completed three measures of anxiety: The Library Anxiety Scale (LAS; Bostick, 1992); the Statistics Anxiety Rating Scale (STARS; Cruise & Wilkins, 1980); and the Composition Anxiety Scale (CARS), a modification of Writing Apprehension Test (Daly & Miller, 1975). These scales measured levels of library anxiety, statistics anxiety, and composition anxiety, respectively. These scores were correlated with the scores that the students attained for their research proposals (i.e., measure of achievement). An *all possible subsets* regression analysis (Onwuegbuzie & Daniel, 2003) revealed that various components of library anxiety, statistics anxiety, and composition anxiety were statistically significant predictors of students' scores on their research proposals. The regression model indicated that students who received low scores in their research proposals tended to have high levels of library anxiety, statistics anxiety, and composition anxiety. The total proportion of variance in achievement scores explained, 35.9%, suggested that these anxiety measures were strong predictors of students' ability to write research proposals (Cohen, 1988), indicating practical significance in addition to the statistical significance found.

The qualitative portion of the research (i.e., Study 2) involved the analysis of these students' reflexive journals using a phenomenological mode of inquiry (Goetz & LeCompte, 1984). The method of constant comparison (Lincoln & Guba, 1985) revealed that the anxiety associated with writing a research proposal represented a multidimensional construct, representing dimensions that included library anxiety, statistics anxiety, and composition anxiety. Students with the highest levels of anxiety in one or more of these areas also tended to be those who reported behaviors (e.g., avoidance behaviors, procrastination) that affected their abilities to write research proposals. These findings not only supported the relationship found in Study 1, but it also helped to explain why this relationship was strong. Thus, the qualitative analysis enhanced the researcher's understanding (i.e., increased *Verstehen)* of the role that anxiety plays in the research proposal writing process.

Concurrent mixed analysis

Concurrent mixed analyses involve the analysis of quantitative and qualitative data types within the same analytical framework. More specifically, in concurrent mixed analyses, quantitative and qualitative data are collected at the same time, and the data analysis typically occurs *after* all the data (i.e., both quantitative and qualitative data) have been collected. However, unlike the case for parallel mixed analyses, integration usually occurs at the data analysis stage. Teaching Evaluation Forms given out to students at the end of the semester at virtually every institution of higher education provide a very common example of how such an analysis can lead to significant findings being enhanced. Typically, these forms extract both quantitative and qualitative information concurrently. The quantitative section of these evaluation forms, which usually carries the most weight, typically is represented by a Likert-format scale consisting of items that assess some aspect of teaching (Onwuegbuzie, Daniel, & Collins, 2004). Responses to these items are then averaged to produce a mean teaching performance score. This average is then used as an index of teaching effect (i.e., effect size). The qualitative section of these forms routinely contains one or more open-ended items asking the respondent to discuss their perceptions of the quality of teaching received. These open-ended responses are then compared to enhance the interpretation of the teaching effect size.

Concurrent mixed analyses also can be undertaken in quantitative studies by qualitizing data, which is a process by which quantitative data are transformed into data that can be analyzed qualitatively (Tashakkori & Teddlie, 1998). For example, Teddlie and Stringfield (1993) conducted a longitudinal study of eight matched pairs of schools initially categorized as being either effective or ineffective with respect to baseline data. Five years after the study begun, these evaluators used eight empirical criteria to re-classify the schools' effectiveness status. These criteria were: (a) norm-referenced test scores, (b) criterion-referenced test scores, (c) time-on-task in classrooms, (d) scores on quality of classroom instruction measures, (e) faculty stability, (f) student attendance, (g) changes in socioeconomic status (SES) of the schools= student bodies, and (h) measures of school "climate." Teddlie and Stringfield converted these quantitative data (i.e., qualitized them) into the following four qualitatively-defined school profiles: (a) stable more effective, (b) stable less effective, (c) improving, and (d) declining. These school profiles were then used to add more meaning to the investigators' longitudinal, evolving perspectives on the schools.

Sequential mixed analysis

In sequential mixed analyses, "multiple approaches to data collection, analysis, and inference are employed in a sequence of phases" (Tashakkori & Teddlie, 1998, pp. 149-150). Here, the data analysis always begins *before* all the data are collected. When the qualitative data analysis stage follows the quantitative data analysis stage, this is called a sequential quantitative-qualitative analysis (Onwuegbuzie & Teddlie, 2003). This involves "forming groups of peoples/settings on the initial basis of [quantitative] data and then comparing the groups on [qualitative] data (subsequently collected or available)" (Tashakkori & Teddlie, 1998, p. 135). The sequential quantitative-qualitative analysis

techniques that can enhance significant results include those identified by Onwuegbuzie and Teddlie (2003): (a) qualitative contrasting case analysis, (b) qualitative residual analysis, (c) qualitative follow-up interaction analyses, and (d) qualitative internal replication analysis.

According to Onwuegbuzie and Teddlie (2003), qualitative contrasting analysis involves first undertaking a quantitative descriptive (i.e., non-inferential) data analysis (e.g., total, *z*-score) on some construct (e.g., self-esteem, perfectionism), and then identifying a proportion (e.g., 20%) or a specific number of those who obtained the lowest and highest scores on the numerical measure. In the second phase, new qualitative data (e.g., interviews, focus groups, observations) are collected on the lowest- and highest-scoring groups, followed by a qualitative analysis (e.g., thematic analysis) of the newly collected data, in an attempt to determine why the two groups differed on the quantitative measure. An example of this is Sheumaker (2001), who, using an instrument she developed called the Technology and Teaching Practices Survey, identified teachers and administrators with the lowest and highest levels of constructivist beliefs. These selectees were then interviewed to determine: (a) what qualitative factors led to their beliefs being so extreme and (b) whether these beliefs stemmed from a constructivist-based staff development program for technology training (i.e., *InTech*) in the state of Georgia.

Qualitative residual analysis involves conducting a General Linear Model (GLM) analysis (e.g., multiple regression), followed by a residual analysis on the selected model in order to identify any outliers (i.e., participants who do not fit the model). In the second phase, new qualitative data are collected on participants who represent the outlying cases, followed by a qualitative analysis (e.g., thematic analysis) of the newly collected data, in an attempt to determine why these participants did not fit the chosen model. For instance, several evaluators (e.g., Kochan, Tashakkori, & Teddlie, 1996; Teddlie & Stringfield, 1993) have classified schools into ineffective and effective groups based on the residual scores from a multiple regression analysis of standardized test scores.

According to Onwuegbuzie (2003b), many researchers neglect to assess the presence of interactions when testing hypotheses. By not formally testing for interactions, researchers may end up selecting a model that does not honor optimally the nature of reality that they want to study, thereby threatening the internal validity of the findings. Because many analysts do not disaggregate their data, they often incorrectly assume that their findings are invariant across all sub-samples inherent in their study. Onwuegbuzie (2003b) termed this "*non-interaction seeking bias*." Therefore, as recommended by Onwuegbuzie, whenever possible, researchers should utilize *condition-seeking* methods, whereby they "seek to discover which, of the many conditions that were confounded together in procedures that have obtained a finding, are indeed necessary or sufficient" (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986, p. 223). By implementing condition-seeking methods, a progression of qualifying conditions are made based on existing findings, which generate a progression of research questions, which, if addressed in future studies, would provide increasingly reliable and generalizable conclusions (Greenwald et al., 1986). Qualitative follow-up interaction analyses are consistent with condition-seeking methods. An example of a qualitative follow-up interaction analyses are studies by Teddlie and Stringfield (1985, 1993). These researchers conducted what was termed a "contextually sensitive" school effectiveness research study, in which

comparisons were made between effective schools that served students from lower-SES environments and those that served students from middle-SES environments. Differences were found that led to the refutation that the correlates of effective schools were generalizable across different school contexts.

Qualitative internal replication analyses involve undertaking a General Linear Model analysis (e.g., multiple regression), followed by an internal replication analysis on the selected model (e.g., jackknife analysis) in order to determine *internal replication outliers* (i.e., cases who unduly affect the internal replication analysis). In the second phase, new qualitative data are collected on those who have been identified as outliers, followed by a qualitative analysis (e.g., thematic analysis) of the newly collected data, in an attempt to determine why these individuals did not fit the chosen model.

Enhancing the Interpretation of Significant Findings in Qualitative Research

The collection, analysis, and interpretation of quantitative data can play an important role in enhancing meaning in qualitative studies. The inclusion of qualitative information can be undertaken either concurrently or sequentially. As is the case for quantitative studies, an array of parallel, concurrent, and sequential mixed methods data analysis techniques can be used to supplement qualitative data analyses.

Parallel mixed analysis

As is the case for parallel mixed analyses in quantitative studies, parallel mixed analysis in qualitative studies involve mixing the qualitative and quantitative data at the interpretation stage of the research process. However, in the former case, the quantitative component of the study is given the most weight, whereas in the latter case, the qualitative component is given priority. The evaluation research undertaken by Senne and Rikard (2002) provides an example of this latter case, namely, parallel mixed analyses in qualitative studies. These researchers undertook a comparative analysis of two PETE portfolio models (curricular interventions during the student teacher experience) to determine their effects on intern perceptions of the utility of the teaching portfolio and intern professional growth. Both quantitative and qualitative data were collected in this study. The qualitative phase of the evaluation, the component with the most weight, involved the interns recording their 15-week teaching experiences in weekly reflection logs. Further, the interns were asked to complete an 8-item questionnaire. This questionnaire was designed for interns to evaluate the portfolio process, the teacher education program, and the student teaching experience. This instrument also asked the interns to describe their accomplishments and overall professional growth. The quantitative phase of the evaluation, which took place concurrently, involved administering a measure of developmental growth (i.e., principled thinking and moral judgment reasoning). The quantitative and qualitative data were analyzed separately before being compared. No statistically significant difference in gain scores in principled thinking and moral judgment reasoning were found for either group of interns. Also, no statistically significant difference in gain scores in principled thinking and moral judgment reasoning was found when both schools were compared. However, the researchers indicated that the statistically non-significant differences likely were the

result of low statistical power resulting from the fact that 30% of one group of interns and 49% of the other group did not complete the quantitative measure. Although no statistically significant evidence of professional growth emerged from the quantitative data, a more positive picture emerged from the qualitative analysis. Specifically, approximately 50% of the interns stated that they became more prepared, assertive, mature, and confident over the course of the internship initiative. Thus, the qualitative data analysis increased the researchers' *verstehen* by supporting their conclusion from the quantitative data analysis that the low statistical power prevented the developmental growth of the interns from being identified via the dependent *t*-test used.

Concurrent mixed analysis

As before, concurrent mixed analyses involve the analysis of quantitative and qualitative data types within the same analytical framework, with integration usually occurring at the data analysis stage. The most common way of supplementing qualitative analysis with a quantitative analysis is by quantitizing data. Quantitizing involves the transformation of the qualitative data to a numerical form (Tashakkori & Teddlie, 1998). More specifically, in quantitizing, "qualitative 'themes' are numerically represented, in scores, scales, or clusters, in order more fully to describe and/or interpret a target phenomenon" (Sandelowski, 2001, p. 231). Witcher, Onwuegbuzie, Collins, Filer, and Wiedmaier (2003) illustrated how emergent qualitative themes could first be quantitized and then subjected to statistical analysis. These researchers examined students' perceptions of characteristics of effective college teachers among 912 undergraduate and graduate students from various academic majors enrolled at a university in a mid-southern state. A qualitative analysis revealed nine characteristics that students considered to reflect effective college teaching. These themes were then "binarized" (Onwuegbuzie, 2003a). Specifically, for each study participant, a score of a "1" was given for a theme if it represented a significant statement or observation pertaining to that individual; otherwise, a score of "0" was given. That is, for each sample member, each theme was *binarized* to a score of "1" or "0." This binarization led to the formation of an *inter-respondent matrix* (*participant x theme matrix*). The inter-respondent matrix indicated which individuals contributed to each emerging theme. This matrix allowed inferential statistical analyses to be conducted. For example, Witcher et al. used a series of Fisher's Exact tests to correlate each of the nine themes with each of the following four demographic variables: gender, race (Caucasian-American vs. minority), level of student (undergraduate vs. graduate), and preservice teacher status (i.e., preservice teacher vs. non-preservice teacher). One result stemming from these chi-square analyses was that females (62.3%) tended to place statistically significantly more weight on student-centeredness as a measure of instructional effectiveness than did males (49.4%). The effect size associated with this relationship, as measured by Cramer's *V*, was .12. Further, females were 1.70 times (95% confidence interval [CI] = 1.26, 2.29) more likely than were males to endorse student-centeredness. With respect to race, Caucasian-American students (31.6%) were statistically significantly more likely to endorse enthusiastic about teaching as a characteristic of effective instruction than were minority students (19.5%). Cramer's *V* effective size was .09. More specifically, Caucasian-American students were 1.61 times (95% CI = 1.12, 2.32) more likely than were minority

students to endorse being enthusiastic about teaching. Several other findings emerged from this series of analyses. Thus, subjecting quantitized data to statistical analysis aided Witcher et al. in the interpretation of the qualitative themes.

Quantitizing data allows qualitative researchers to enhance meaning further by reporting effect sizes associated with qualitative observations (Onwuegbuzie, 2003a). In its simplest from, effect sizes in qualitative research represent counts of observations or themes. Building on Onwuegbuzie's (2003a) conceptualization, Sandelowski and Barroso (2003) stated the following:

> Although qualitative studies typically do not address treatments, they do address patterns and themes, which inherently imply a frequency of occurrence of an even sufficient to constitute a pattern on theme….The calculation of effect sizes constitutes a quantitative transformation of qualitative data in the service of extracting more meaning from those data and verifying the presence of a pattern or theme. Effect sizes in qualitative studies are both a means to ensure that findings are neither over- nor under-weighted, and the final form in which a metasummary of findings might appear. (p. 231)

Sechrest and Sidani (1995, p. 79) note that, "qualitative researchers regularly use terms such as 'many,' 'most,' 'frequently,' 'several,' 'never,' and so on. These terms are fundamentally quantitative." Thus, qualitative researchers can obtain more meaning by obtaining counts of observations in addition to their narrative descriptions (Sandelowski, 2001). For instance, Witcher et al. (2003), using the inter-respondent matrix described above, counted the frequency of the emergent themes. These researchers found that of the nine identified characteristics of effective college teachers, student-centeredness was the most commonly-cited trait (cited by 58.9% of the sample). This was followed by knowledge of subject matter (44.1%), professionalism (40.8%), enthusiasm about teaching (29.8%), effective communication (23.5%), accessibility (23.3%), competent instruction (21.8%), fairness and respectfulness (21.6%), and provider of adequate performance feedback (5.0%). Providing these prevalence rates enhanced *verstehen* by preventing the researcher from over-weighting or under-weighting the emergent themes (Sandelowski, 2001).

Miles and Huberman (1994) contend that the identification of categories, codes, themes, typologies, and the like are based, at least to some extent, on the frequency with which a phenomenon occurs (Miles & Huberman, 1994). According to these authors, there are three reasons for counting themes: (a) to identify patterns more easily, (b) to verify a hypothesis, and (c) to maintain analytic integrity. Further, by adding numerical precision to their descriptive narratives, Witcher et al. (2003) were able to leave an audit trail, which are recommended by many qualitative researchers as a method of evaluating legitimation or increasing legitimation, or both (Halpern, 1983; Lincoln & Guba, 1985).

Onwuegbuzie and Teddlie (2003) provided a typology for reporting effect sizes alongside qualitative observations. These indices comprise "manifest effect sizes" and "latent effect sizes." Manifest effect sizes are effect sizes that quantify observable content. According to Onwuegbuzie and Teddlie (2003, p. 356), "this class of effect sizes represents specific counts (or percentages) of significant statements (e.g., words, phrases,

sentences, paragraphs, pages) or observations analyzed that underlie emergent themes." Manifest effect sizes can be further subdivided into frequency manifest effect sizes and intensity manifest effect sizes. Frequency manifest effect sizes determine the prevalence rates of themes or observations. For example, the prevalence rates of the perceived characteristics of effective college instructors of Witcher et al. (2003) documented above represent frequency manifest effect sizes. Intensity manifest effect sizes represent "the frequency of each significant statement within each theme, or the frequency of each theme within a set of themes" (p. 356). Effect sizes can be adjusted for the length of the unit of analysis (e.g., observation, text, interview). For instance, the prevalence of a theme can be divided by the number of transcribed words, sentences, paragraphs, or pages analyzed, yielding an adjusted effect size. Also, a fixed-interval effect size index could be determined, wherein "the frequency (i.e., fixed-interval frequency effect size) and intensity (i.e., fixed-interval intensity effect size) of themes are determined as they occur within a specific period of time. For example, a researcher could investigate how many times a word is used in the first 10 minutes of a focus group" (Onwuegbuzie & Teddlie, 2003, p. 358). In addition, a *fixed-ratio effect size index* could be estimated, wherein "a specific frequency (i.e., fixed-response frequency effect size) and intensity (i.e., fixed-response intensity effect size) of themes are specified *a priori*, and the amount of time that elapses before these targets are met, if at all, is utilized as an effect size estimate" (Onwuegbuzie & Teddlie, 2003, p. 358).

Latent effect sizes also could be used to provide more meaning to qualitative observations. These indices, in contrast to manifest effect sizes, represent effect sizes that quantify non-observable content. For example, Witcher et al. (2003) undertook a canonical correlation analysis to examine the multivariate relationship between the nine themes presented above and eight demographic variables (gender, race, level of student, student teacher status, age, GPA, number of credit hours taken, and number of offspring). The first canonical correlation indicated that gender, level of student, preservice teacher status, and number of credit hours related to student-centeredness, professionalism, fairness and respectfulness, and competent instructor. This first canonical correlation (R_{c1} = .31) was deemed by Witcher et al. to be moderately practically significant, contributing 9.6% (i.e., R_{c1}^2) to the shared variance. The proportion of shared variance, R_{c1}^2, served as a (variance-explained) latent effect size, which provided much more understanding about students' perceptions of the characteristics of effective college instructors than would have been obtained if only a qualitative (i.e., thematic) analysis had been undertaken.

Sandelowski and Barroso (2003) showed how effect sizes can be used to conduct metasummaries of qualitative findings. According to these methodologists, a qualitative metasummary is "a form of systematic review or integration of qualitative findings in a target domain that are themselves topical or thematic summaries or surveys of data" (p. 227). They conducted a qualitative metasummary of 45 published and unpublished reports of qualitative studies of HIV-positive women with results on motherhood, which led to 800 findings being extracted, which were reduced to 93 abstracted findings, from which manifest frequency and intensity effect sizes were calculated. Sandelowski and Barroso found that five results had effect sizes ranging from 25% to 60%, with both published and unpublished articles contributing approximately equally to the strength of these findings. A total of 73 findings had effect sizes that were less than 9%, with 47 of them having effect sizes of only 2%.

Sequential mixed analysis

In sequential qualitative-quantitative analysis, an initial qualitative data analysis leads to the identification of groups of individuals who are similar in some respect to each other. These identified groups are then compared to each other using either existing quantitative data, or data that are collected after the initial qualitative data analysis (Onwuegbuzie & Teddlie, 2003). For example, Daley and Onwuegbuzie (2004) investigated male juvenile delinquents' causal attributions for others' violent behavior, and the salient pieces of information they utilize in arriving at these attributions. They developed an instrument that they called the Violence Attribution Survey, a 12-item questionnaire designed to assess attributions made by juveniles for the behavior of others involved in violent acts. Each item on this instrument consisted of a vignette, followed by three possible attributions (i.e., person, stimulus, circumstance) presented in multiple-choice format, and an open-ended question asking the juveniles their reasons for choosing the response that they did. Eighty-two male juvenile offenders, selected via an *a priori* power analysis, were involved in this study. These offenders were drawn randomly from the population of juveniles incarcerated at a correctional facility in a large southeastern state. A phenomenological analysis revealed the following seven themes that arose from juveniles' reasons for their causal attributions: self-control, violation of rights, provocation, irresponsibility, poor judgment, fate, and conflict resolution. Daley and Onwuegbuzie conducted an ipsative/cluster analysis on these themes, and identified three distinct profiles of delinquents. These three profiles of delinquents were compared on a number of quantitative measures, with age differences emerging.

Onwuegbuzie and Teddlie (2003) have identified the following types of sequential qualitative-quantitative analyses: (a) quantitative extreme case analysis and (b) quantitative negative case analysis. Quantitative extreme case analysis involves first undertaking a qualitative data analysis (e.g., thematic analysis). This qualitative analysis is then followed by a legitimation analysis in order to determine the extreme cases. In the second phase, new quantitative data are collected on all cases, followed by a quantitative analysis (e.g., *t*-test) of the newly collected quantitative data, wherein the extreme and non-extreme cases are compared, in an attempt to determine why the former cases were so extreme in the first phase.

Quantitative negative case analysis involves conducting a qualitative data analysis (e.g., thematic analysis), followed by a legitimation analysis, in order to identify negative cases (i.e., participants who do not fit the interpretation or initial theory). In the second phase, new quantitative data are collected on all cases, followed by a quantitative analysis (e.g., *t*-test) of the newly collected data, in which the negative and non-negative cases are compared, in an attempt to determine why the former did not fit the model in the first phase.

Summary and Conclusions

The goal of the present paper was to outline how mixed methods research can be used to enhance the interpretation of *significant* findings in educational evaluation research studies. First, we defined what we mean by significance in educational evaluation research. With regard to quantitative-based research, we defined the four types of significance: statistical significance, practical significance, clinical significance, and

economic significance. With respect to qualitative-based research, we defined a significant finding as one that achieves understanding or *verstehen*. Second, we presented some of the major limitations of each of these types of significance. In particular, we contended that all of these types of significance provide partial information at best, and that sole reliance on any of these indices can lead to misleading interpretations of the data. Third, we described how conducting mixed methods analyses can be used to enhance the interpretation of significant findings in both quantitative and qualitative evaluations. With respect to quantitative-based evaluations, we demonstrated how the collection, analysis, and interpretation of qualitative data could aid the interpretation of statistically significant, practically significant, clinically significant, and economically significant findings. With regard to qualitative findings, we illustrated how quantitative data collection, analysis, and interpretation could add meaning.

Interestingly, every method for enhancing the interpretation of significant findings described in this treatise is compatible with one of more of Greene et al.'s (1989) five purposes of mixed methods research (triangulation, complementarity, development, initiation, expansion). In particular, conducting a parallel mixed analysis either in a predominantly quantitative or qualitative study is consistent with the goals of triangulation, complementarity, and initiation. A concurrent mixed analysis either in a primarily quantitative or qualitative investigation can be used for the purposes of triangulation, complementarity, development, and initiation. Finally, utilizing a sequential mixed analysis either in a predominantly quantitative or qualitative investigation can address the goals of complementarity, development, and expansion. In any case, we contend that conducting mixed methods analyses in a parallel, concurrent, or sequential manner is more likely to lead to ethical research outcomes in both quantitative and qualitative evaluation studies. As such, use of mixed methods data-analytical techniques should be seen as the real gold standard for achieving *verstehen* in educational evaluation research.

References

Anfara, V. A., Brown, K. M., & Mangione, T. L. (2002). Qualitative analysis on stage: Making the research process more public. *Educational Researcher, 31*(7), 28-38.

Bogdan, R. C., & Biklen, S. K. (2003). *Qualitative research for education: An introduction to theories and methods*. Boston: Pearson Education Group.

Bostick, S. L. (1992). The development and validation of the library anxiety scale. *Dissertation Abstracts International, 53* (12), 4116A. (Publication No. AAT9310624)

Carspecken, P. F., & Apple, M. (1992). Critical qualitative research: Theory, methodology, and practice. In M. D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 507-553). San Diego, CA: Academic Press

Chambers, E. (2000). Applied ethnography. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 851-869). Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1997). The earth is round (p < .05). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

Connolly, P. (1998). 'Dancing to the wrong tune': Ethnography generalization and research on racism in schools. In P. Connolly & B. Troyna (Eds.), *Researching racism in education: Politics, theory, and practice* (pp. 122-139). Buckingham, UK: Open University Press.

Constas, M. A. (1992). Qualitative data analysis as a public event: The documentation of category development procedures. *American Educational Research Journal, 29,* 253-266.

Cruise, R. J., & Wilkins, E. M. (1980). *STARS: Statistical Anxiety Rating Scale*. Unpublished manuscript, Andrews University, Berrien Springs, MI.

Daley, C. E., & Onwuegbuzie, A. J. (2004). Attributions toward violence of male juvenile delinquents: A concurrent mixed methods analysis. *Journal of Social Psychology, 144*(6), 549-570.

Daly, J. A., & Miller, M. D. (1975). The empirical development of an instrument to measure writing apprehension. *Research in the Teaching of English,* 9, 242-249.

Deyhle, D. L., Hess, A., Jr., & LeCompte, M. D. (1992). Approaching ethical issues for qualitative researchers in education. In M. D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 595-641). San Diego, CA: Academic Press.

Denzin. N. K., & Lincoln, Y. S. (2000). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1-28). Thousand Oaks, CA: Sage.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94,* 275-282.

Gay, L. R., & Airasian, P. W. (2003). *Educational research: Competencies for analysis and application* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Goetz, J. P., & LeCompte, M. D. (1984). *Ethnography and the qualitative design in educational research.* New York: Academic Press.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255-274.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress. *Psychological Review, 93,* 216-229.

Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology, 29*(2), 75-91.

Halpern, E. S. (1983). *Auditing naturalistic inquiries: The development and application of a model*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60,* 543-563.

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 297-319). Thousand Oaks, CA: Sage.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332-339.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*(3), 285-296.

Kirk, R. E. (1996). Practical significance. A concept whose time as come. *Education and Psychological Measurement, 56*, 746-759.

Kochan, S., Tashakkori, A., & Teddlie, C. (1996, April). *You can't judge a high school by achievement alone: Preliminary findings from the construction of behavioral indicators of school effectiveness*. Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

LeCompte, M. D. (1987). Bias in the biography: Bias and subjectivity in ethnographic research. *Anthropology and Education Quarterly, 18*, 43-52.

Leech, N. L., & Onwuegbuzie, A. J. (in press). A proposed fourth measure of significance: The role of economic significance in educational research. *Evaluation and Research in Education.*

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Madey, D. L. (1982). Some benefits of integrating qualitative and quantitative methods in program evaluation, with some illustrations. *Educational Evaluation and Policy Analysis, 4*, 223-236.

Majova-Seane, N. (2003, February). *The inclusion of effect sizes in addition to statistical significance testing reporting*. Paper presented at the annual meeting of the Southwestern Educational Research Association, San Antonio, TX.

Miles, M., & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Newman, I., Ridenour, C. S., Newman, C., & DeMarco, G. M. P. (2003). A typology of research purposes and its relationship to mixed methods. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 167-188). Thousand Oaks, CA: Sage.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286.

Onwuegbuzie, A. J. (1997). Writing a research proposal: The role of library anxiety, statistics anxiety, and composition anxiety. *Library & Information Science Research, 19*, 5-33.

Onwuegbuzie, A. J. (2003a). Effect sizes in qualitative research: A prolegomenon. *Quality & Quantity: International Journal of Methodology, 37*, 393-409.

Onwuegbuzie, A. J. (2003b). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools, 10*(1), 71-90.

Onwuegbuzie, A. J., & Daniel, L. G. (2003, February 12). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education, 6*(2). Retrieved on February 18, 2005, from http://cie.ed.asu.edu/volume6/number2/

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2004, February). *Problems associated with student teacher evaluations.* Paper to be presented at the annual meeting of the Eastern Educational Research Association, Clearwater, FL.

Onwuegbuzie, A. J., & Johnson, R. B. (2004). Mixed research. In R. B. Johnson & L. B. Christensen (Eds.), *Educational research: Quantitative, qualitative, and mixed approaches* (2nd ed., pp. 408-431). Needham Heights, MA: Allyn & Bacon.

Onwuegbuzie, A. J., & Leech, N. L. (in press). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology: Theory & Practice.*

Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods, 2*, 133-151.

Onwuegbuzie, A. J., Levin, J. R., & Leech, N. L. (2003). Do effect-size measures measure up? A brief assessment. *Learning Disabilities: A Contemporary Journal, 1,* 37-40

Onwuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351-383). Thousand Oaks, CA: Sage.

Outhwaite, W. (1975). *Understanding social life: The method called Verstehen*. London: Allen & Unwin.

Patton, M. Q. (1990). *Qualitative evaluation methods* (2nd ed.). Thousand Oaks, CA: Sage.

Richardson, L. (2000). Writing: A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 923-948). Thousand Oaks, CA: Sage.

Rossman, G. B., & Wilson, B. L. (1985). Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review, 9*, 627-643.

Sandelowski, M. (2001). Real qualitative researchers don't count: The use of numbers in qualitative research. *Research in Nursing & Health, 24*, 230-240.

Sandelowski, M., & Barroso, J. (2003). Creating metasummaries of qualitative findings. *Nursing Research, 52*, 226-233.

Schwandt, T. A. (2000). Three epistemological stances for qualitative inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 189-213). Thousand Oaks, CA: Sage.

Schwandt, T. A. (2001). *Dictionary of qualitative inquiry* (2nd ed.). Thousand Oaks, CA: Sage.

Sechrest, L., & Sidana, S. (1995). Quantitative and qualitative methods: Is there an alternative? *Evaluation and Program Planning, 18*, 77-87.

Senne, T. A., & Rikard, G. L. (2002). Experiencing the portfolio process during the internship: A comparative analysis of two PETE portfolio models. *Journal of Teaching in Physical Education, 21*, 309-336.

Sheumaker, M. F. (2001). *Technology integration in middle school classrooms: The influence of staff development on the beliefs of teachers and administrators.* Unpublished doctoral dissertation, Valdosta State University, GA.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches.* Applied Social Research Methods Series (Vol. 46). Thousand Oaks, CA: Sage.

Teddlie, C., & Stringfield, S. (1985). A differential analysis of effectiveness in middle and lower socioeconomic status schools. *Journal of Classroom Interaction, 20*(2), 38-44.

Teddlie, C., & Stringfield, S. (1993). *Schools make a difference: Lessons learned from a 10-year study of school effects.* New York: Teachers College Press.

Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*, 361-377.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development, 80*, 64-71.

Vacha-Haase, T. (2001, April). *Addressing the clinical significance as against statistical or practical significance: science and practice working together to benefit the practitioner.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Vaske, J. J., Gliner, J. A., & Morgan, G. A. (2002). Communicating judgments about practical significance: Effect sizes, confidence intervals, and odds ratios. *Human Dimensions of Wildlife, 7*, 287-300.

Willems, E. P., & Raush, H. L. (1969). *Naturalistic viewpoints in psychological research.* New York: Holt, Rinehart, & Winston.

Witcher, A. E., Onwuegbuzie, A. J., Collins, K. M. T., Filer, J., & Wiedmaier, C. (2003, November). *Students' perceptions of characteristics of effective college teachers.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS.

Author Note

Anthony J. Onwuegbuzie (Ph.D., University of South Carolina) is associate professor in the Department of Educational Measurement and Research at the University of South Florida. He teaches courses in qualitative research, quantitative research, and mixed methods. His research topics primarily involve disadvantaged and under-served populations such as minorities and juvenile delinquents. He also writes extensively on quantitative, qualitative, and mixed methodological topics.

Nancy L. Leech (Ph.D., Colorado State University) is assistant professor in the Division of Educational Psychology at the University of Colorado at Denver and Health Sciences Center. She teaches courses in qualitative research and quantitative research. Her research topics include understanding success in higher education, gender and equity issues, and various methodological concepts and techniques.

Correspondence should be addressed to Anthony J. Onwuegbuzie, Department of Educational Measurement and Research, College of Education, University of South Florida, 4202 East Fowler Avenue, EDU 162, Tampa, FL, 33620-7750, or E-Mail: (tonyonwuegbuzie@aol.com)

Author citation

Onwuegbuzie, A. J., Leech, N. L. (2004). Enhancing the interpretation of "significant" findings: The role of mixed methods research. *The Qualitative Report, 9*(4), 770-792. Retrieved [Insert date], from http://www.nova.edu/ssss/QR/QR9-4/onwuegbuzie.pdf